ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

A Survey of Phishing Website Detection and Prevention Techniques

Parasmani M Rathod [1], Arnab Gajbhiye [1] , Rahul Atri [1], Anita Sachin Mahajan [2]

U.G. Student, Department of Computer Engineering, Dr D.Y Patil School of Engg. and Technology, Lohgaon, Pune, Maharashtra, India [1]

Assistant Professor, Department of Computer Engineering, Dr D.Y Patil School of Engg. and Technology, Lohgaon, Pune, Maharashtra, India [2]

ABSTRACT: Phishing attack has become one of the most serious issues in today's world. Many people have lost huge sums of money by becoming victim to these attacks. Phishing websites trick internet users into divulging their personal or financial information such as username, password, credit card details, bank details etc. These websites look exactly like the original website. Various anti-phishing techniques make use of different features of the webpage in order to identify the fraud website. In this paper, we discuss different phishing detection techniques and present the survey of various phishing website detection approaches.

KEYWORDS: phishing, data mining, SVM, machine learning, legitimate website, phishing website

I. INTRODUCTION

Phishing is a cybersecurity threat in which phishers trick internet users into divulging their confidential and personal information. Phishers also use some social engineering techniques to steal user data. The phishing attack can happen in many ways such as email, website and through a computer virus. In most phishing attacks emails are sent to the user. These emails pretend to be from a legitimate organisation and redirect the user to a website wherein user enters his or her personal and confidential information such as username, password, credit card details, bank details etc. These websites also called as phishing website now steal the user information. Phishers may sell this information or use it to carry out illegal tasks to obtain personal gain.

Phishing attacks are mainly classified into four categories such as Phishing, Spear phishing, Clone phishing and Whaling.

1. Phishing: In this technique, the attacker pretends to be the legitimate entity in order to obtain personal and confidential information such as username, password, credit card details, bank account details etc.
2. Spear Phishing: Unlike ordinary phishing attack in which the phishing emails are sent to the multiple users, in spear phishing attack a specific organisation or individual is targeted. The emails appear to come from a known recipient.
3. Clone Phishing: In this technique, the phishy email account mimics the original email address. The links and attachments are usually replaced with a malicious version.
4. Whaling: This phishing technique targets popular people such as senior executives, celebrities, politicians, business tycoons etc. The main aim of this attack is to swindle someone into revealing confidential company or institute information.

As per the APWG (Anti-Phishing Working Group) report [10] for Q4 2016, the total number of phishing attacks reported by consumers in 2016 was 1,220,523 which is a 65 percent increase over 2015. This significant increase in phishing attacks shows the severity of this problem.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

II. LITERATURE SURVEY

A. CANTINA

Zhang et al [1] proposed a content-based phishing detection technique called CANTINA. CANTINA uses TF-IDF short for term frequency-inverse document frequency information retrieval algorithm. TF provides most frequently occurring terms in the document and IDF gives the measure of how unique the word is i.e. how infrequently the word occurs across all documents.
CANTINA works in the following steps:
1. Calculate TF-IDF of each term or word present on that web page.
2. Find top five terms with highest TF-IDF weight.
3. Generate lexical signature of these terms.
4. Provide these signatures as input to search engine.
5. Compare the domain name of the web page with top results provided by the search engine.

B. CANTINA+

Xiang et al [2] proposed a system CANTINA+ which is a modified and upgraded version of CANTINA. CANTINA+ uses eight new features which makes use of HTML, Document Object Model, search engines and few third-party services. Unlike CANTINA it is evaluated on a larger set of URLs. It consists of two filters which help to reduce false positive and to increase the speed of the system. It extracts fifteen features from each URL and arranges them in a proper format. After that, it applies machine learning techniques and builds the classifiers to detect phishing website. It checks for the availability of login form in a web page. If the login form is found then the web page will be sent to the feature extractor otherwise it will be classified as legitimate.

C. ASSOCIATIVE CLASSIFICATION DATA MINING

Neda Abdelhamid et al [3] proposed Multi-Label Classifier based Associative Classification Method. It provides high accuracy in detecting phishing websites. It generates new rules which are used to enhance its classifier predictive performance. MCAC looks for the hidden relation between attribute values and class attributes in the training dataset. It uses these relations to generate association rules. After creating rules it sorts them using different sorting algorithms. It ignores repeated rules and provides accuracy in terms of true positive and false positive.

D. A SVM BASED APPROACH

Huang et al [4] proposed a technique based on URL based features. it extracts 23 features from URL and applies SVM (Support Vector Machine) technique to train the system. The system makes use of lexical and brand name features of URL to take decisions.

E. CLASSIFICATION MINING TECHNIQUES

M. Aburrous et al [5] proposed a technique based on classification data mining techniques. It is designed to detect e-banking phishing websites. In this technique, author implements six different classification algorithms. The different classification algorithms used are PRISM, PART, CBA, C4.5, JRip and MCAR. These algorithms are tested on different training data sets to measure the performance and accuracy of the system. It first generates rules and based on these rules the classifiers are created. The false positive rate is very high therefore this technique is not very efficient.

F. URL-ASSISTED BRAND NAME WEIGHTING SYSTEM

C L Tan et al [6] proposed a technique called URL-assisted brand name weighting system. It primarily focuses on websites with English content. It checks for the position of the brand name in the URL. It exploits this pattern by extracting words from the web page and assigning TF-IDF weights to them. The most weighted words are provided as

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

input to the search engine. A WHOIS lookup is performed to get information about domain name owner. If the information matches then the website is considered as legitimate otherwise fraudulent. This method provides high accuracy with a very low false positive rate.

G. NOVEL APPROACH USING AUTO UPDATED WHITELIST

Ankit Kumar Jain et al [7] proposed an approach to detect phishing attack which is based on the auto-updated whitelist. This system works on mainly two modules namely URL and DNS. DNS module contains two attributes Domain name and its IP address. It checks for the given domain name and its corresponding IP address in the whitelist. If the match is found then the website is considered as legitimate otherwise fraudulent. If the website is not in the whitelist and user is visiting the site for the first time then the examination of the features of the hyperlink is done and an alert is sent to the user if the hyperlink is found to be fake. In case it is found legitimate then the website is added to the whitelist.

H. WEB PHISHING DETECTION BASED ON GRAPH MINING

Zou Futai et al [8] present a technique based on graph mining. It mainly focuses on the behavioural pattern of phishing websites. It is based on ISP's real traffic. The algorithm used in graph mining constructs undirected graph including user nodes and URL nodes according to traffic flow data at first and then through the BP algorithm (Brief Propagation) of MRF (Markov Random Field).After that, it iteratively corrects node reputation and defines reputation threshold to detect phishing. This technique makes use of the inherent relationship between URL and the user.

I. LATENT DIRICHLET ALLOCATION AND ADABOOST

V Ramanathan et al [9] proposed a solution which utilizes Latent Dirichlet Allocation and Adaboost. In this method, different features of the web page are extracted using Latent Dirichlet Allocation and classified using Adaboost. This method is a content-driven approach that is content independent and language neutral. The contents of the website are collected by employing an intelligent web crawler. The contents which are not in English are translated using Google's language translator.

III. **COMPARISON**

TABLE. I ADVANTAGES AND DRAWBACKS OF VARIOUS PHISHING WEBSITE DETECTION TECHNIQUES

Approach	**Advantages**	**Drawbacks**
CANTINA [1]	Fast	Accuracy is completely dependent on TF-IDF algorithm
CANTINA+ [2]	High true positive rate	It fails to detect DNS compromised URLs
Associative classification data mining [3]	Employs MCAC applicability in detecting phishing website	Accuracy of system is unknown or not calculated

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

A SVM based approach [4]	It is content independent hence can detect websites which are made of images	Low accuracy as compared to other techniques
Classification mining techniques [5]	Designed to prevent e-banking frauds	False positive rate is high
URL assisted brand name weighing system [6[	High true positive rate	Since it is URL based hence accuracy is low.
Novel approach using auto updated whitelist [7]	Low false positive rate	High running time
Web Phishing Detection Based on Graph Mining [8]	Supports distributed computing	Running time is high
Phishing website detection using Latent Dirichlet Allocation and Adaboost [9]	Can easily identify zero hour attack	Selection of features is not justified

IV. ISSUES AND CHALLENGES IN FUTURE

We have discussed various phishing detection techniques in detail. There is no single technique that can detect all types of phishing websites. Different parameters are responsible for the accuracy of given technique or solution. It depends on selection of features, the algorithm used and the training data set. Most of the techniques discussed can only be used for the webpages which are in English language. For other languages these techniques cannot be used. e.g Various e-commerce websites are in different regional languages and these techniques may not be able to detect keywords from these websites. Attackers always use different ways to fool internet users, therefore, the phishing detection techniques also need to change as per the phishing trends and environment. Most of the phishing techniques are unable to detect zero hour attack. So these issues need to be solved in future.

V. CONCLUSION

Phishing is one of the most dangerous threats in contemporary world. Victims of these attacks have increased tremendously for last ten years. Cybercriminals change their strategies from time to time in order to gain people's personal information and to crack security system without getting detected. In this paper we have surveyed and discussed various phishing detection techniques. Different techniques use different features of the web page such as URL, text, security certificates, host information etc. in order to detect phishing websites. Techniques discussed in this paper have different drawbacks in terms of accuracy and performance. There is no single system that can detect all types of phishing attacks therefore in future there needs to be an all in one system that will detect all these attacks with high accuracy and performance. Most of the techniques still have limitations in terms of zero hour attack, embedded objects in web pages, computational power, accuracy and performance. All these issues need to be solved in future.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

REFERENCES

[1] Y. Zhang, J. Hong and L. Cranor, "Cantina: a content-based approach to detecting phishing websites", in Proceeding of the 16th international conference on World Wide Web, 2007.
[2] G. Xiang, J. Hong, C. Rose and L. Cranor, "Cantina+: A feature-rich machine learning framework for detecting phishing websites", ACM Transaction on Information and System Security, vol. 14, no. 2, 2011.
[3] Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah, "Phishing detection based Associative classification data mining", Expert systems with Applications, volume 41, Issue 13, pp. 5948-5959, 2014.
[4] H. Huang, L. Qian, Y. Wang, "A SVM-based techniques to detect phishing URLs", Inf.Technol. J. vol. 11, no. 7, pp. 921925, 2012.
[5] M. Aburrous, MA Hossain, K Dahal, T. Fadi, "Predicting phishing websites using classification mining techniques", In: Seventh International conference on information technology, Las Vegas, Nevada, USA, 2010.
[6] C. L. Tan, K. L. Chiew and S. N. Sze, "Phishing website detection using URL-assisted brand name weighting system", 2014 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Kuching, 2014.
[7] Ankit Kumar Jain, B.B. Gupta, "A novel approach to protect against phishing attacks at client side using auto updated whitelist", EURASIP Journal on Information Security (2016) 2016:9 DOI 10.1186/s13635-016-0034-3.
[8] Zou Futai, Gang Yuxiang, Pei Bei, Pan Li, Li Linsen, "Web phishing detection based on graph mining", 2016 Second IEEE International Conference on Computer and Communications, China, 2016.
[9] V. Ramanathan, H. Wechsler, "Phishing website detection using Latent Dirichlet Allocation and AdaBoost", IEEE International conference on Intelligence and Security Informatics. Cyberspace, Border and Immigration Securities, Piscateway, NJ, pp. 102107, 2012.
[10] https://docs.apwg.org/reports/apwg_trends_report_q4_2016.pdf.